

18 | 网络篇：如何步步拆解处理复杂的网络性能问题？

2020-01-06 庄振运

性能工程高手课

[进入课程 >](#)



讲述：庄振运

时长 14:56 大小 11.98M



你好，我是庄振运。

前面几讲，我们讨论了 CPU、内存和存储系统的性能。不过你也清楚，各种互联网服务的数据传递终归是需要通过网络来传输的，所以网络性能也是至关重要的。所以，今天我们就来接着学习网络性能相关的内容。

这一讲我的讲解，依然是会逐步递进，先介绍有关网络的几个性能指标，并从单机的网络性能说起，之后推进到数据中心内部的网络性能，接着再进一步阐释互联网和内容分发网络的性能。最后，再为你介绍几个常用网络性能工具。

网络的性能指标

先来说说网络的性能指标。网络性能五个常用的指标是：

1. 可用性 (availability)
2. 响应时间 (response time)
3. 网络带宽容量 (network bandwidth capacity)
4. 网络吞吐量 (network throughput)
5. 网络利用率 (network utilization)

第一个指标是**可用性**，理解起来比较简单。你肯定有过网络不通（也就是断网）的糟糕体验。没有网络，你几乎啥都干不了。所以，对于网络来讲，最重要的就是，网络是否可以正常联通。

如何测试网络可用性呢？

最简单的方法，就是**使用 ping 命令**。这个命令其实就是向远端的机器发送 ICMP 的请求数据包，并等待接收对方的回复。通过请求和应答返回的对比，来判断远端的机器是否连通，也就是网络是否正常工作。

第二个指标是**响应时间**。端到端的数据一次往返所花费时间，就是响应时间。响应时间受很多因素的影响，比如端到端的物理距离、所经过网络以及负荷、两端主机的负荷等。

第三个指标是**网络带宽容量**。它指的是在网络的两个节点之间的最大可用带宽，这一指标一般是由设备和网络协议决定的，比如网卡、局域网和 TCP/IP 的特性。如果是向网络提供商购买的带宽，那么购买的数量就是网络的带宽容量。

第四个指标**网络吞吐量**是指在某个时刻，在网络中的两个节点之间，端到端的实际传输速度。网络吞吐量取决于当前的网络负载情况，而且是随着时间不同而不断变化的。

第五个指标**网络利用率**是指网络被使用的时间占总时间的比例，一般以百分比来表示。因为数据传输的突发性，所以实际中的网络利用率一般不会太高。否则的话，那么响应时间就不能保证。

单机的网络性能

了解了网络方面性能的指标后，我们接着讨论具体的网络性能。虽然网络传输需要两端进行，但是我们必须从单机开始，来清楚地了解网络的协议栈。

网络协议其实相当复杂，而且分很多层级。操作系统内核中，最大的一个子系统或许就是网络。网络子系统由多个协议组成，其中每个协议都在更原始的协议之上工作。大学时你应该学过 OSI 模型，或 TCP / IP 协议栈，里面的分层定义对网络协议非常适用。当用户数据通过网络协议栈传递时，数据会被封装在该协议的数据包中。

在考虑多层协议的交互时，我们尽量把思路简化。其实它们之间的关系很简单，就是网络协议栈的每一层都有其职责，与其他高层和低层协议无关。

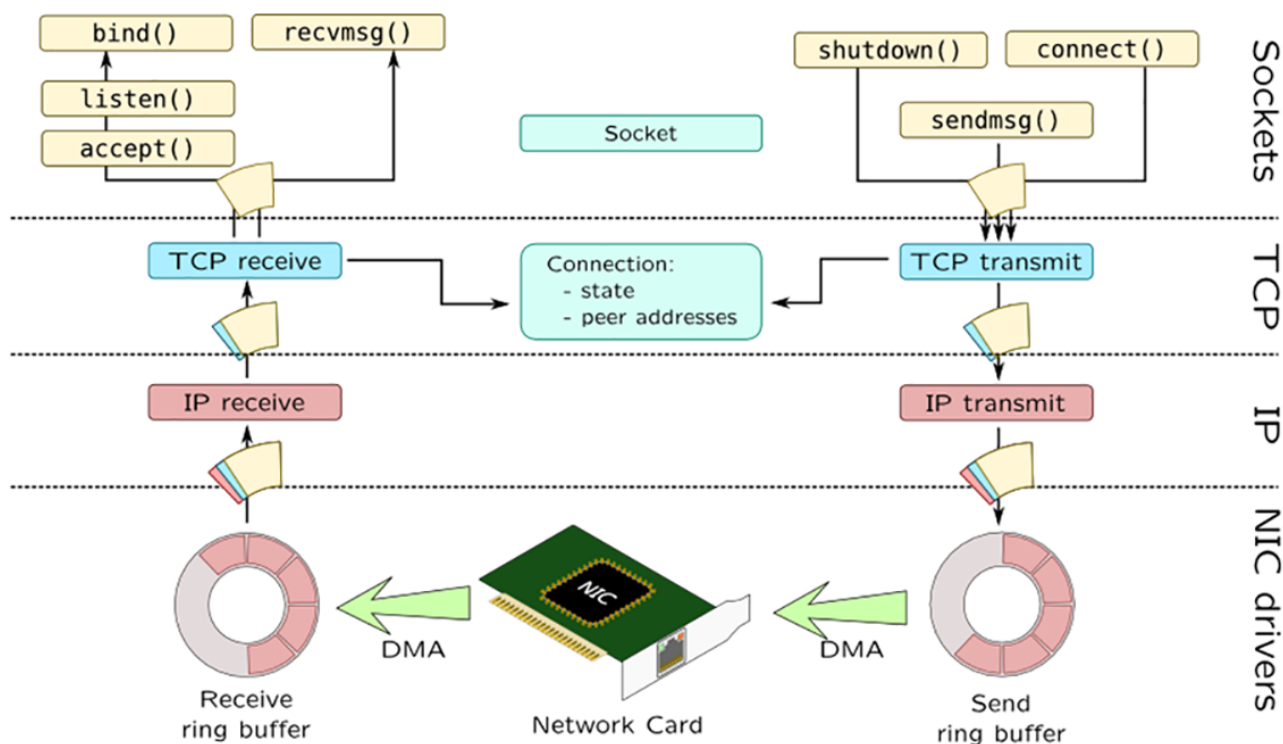
举个例子来说，我们看 IP 层的协议。IP 是第三层协议，它是通过路由器和网络发送端到端的数据报；它的主要目的，就是在网络中的每一段找到路由路径，从而最终能够到达数据报的接收地。但它不保证数据的有序性，也就是转发过程中几个数据报会重新排序；同时也不能保证整个数据的完整性和可靠性，比如丢失的数据报那就是丢失了，IP 层不会重传。

那么数据的可靠性是谁保证的呢？这就需要更上层的 TCP 协议实现。TCP 层协议通过检测数据丢失并且重传，来保证数据的可靠性，也通过序列号来保证数据的有序性。

但是这两层协议：TCP 和 IP 两层协议，都只能传输原始的数据；而对于数据本身是否被压缩过，这些数据表示什么，是在更高层的协议（如 HTTP 和应用层）上实现的。

我们接着具体到网络协议的程序实现，在 Unix 系统中，网络协议栈可以大体分为三层。从上到下，第一层是通过一系列系统调用，实现 BSD 套接字的套接字层，比如 `sendmsg()` 函数；第二层是中间协议的程序，例如 TCP/IP/UDP；第三层是底部的媒体访问控制层，提供对网络接口卡（NIC）本身的访问。

如下图所示（图片来自于 <https://myaut.github.io/>）。



注意这个图，右边是发送端，左边是接收端。先从右边看起，最上层就是 BSD 套接字。它通过一系列 API 调用（比如 connect() 等），然后是 TCP 的发送，再到 IP 的发送，最后到网卡的发送缓冲区，并最终通过网卡发出。接收端，也就是图的左边，则会经过相反的顺序，逐层到达应用层。

数据中心的网络性能

谈完单机的网络性能，我们接着讨论端到端的互联网数据传输。

一台服务器和互联网的远端服务器进行数据传输的时候，需要经过好几层交换器和不同的网络。数据从一台服务器的网卡出来之后，下一步就是经过机柜上面的交换器，然后是数据中心内部的网络，再进入互联网骨干网络。接收端的情况正好相反。

对于这些中间的网络构件，我们一个一个地讨论一下。

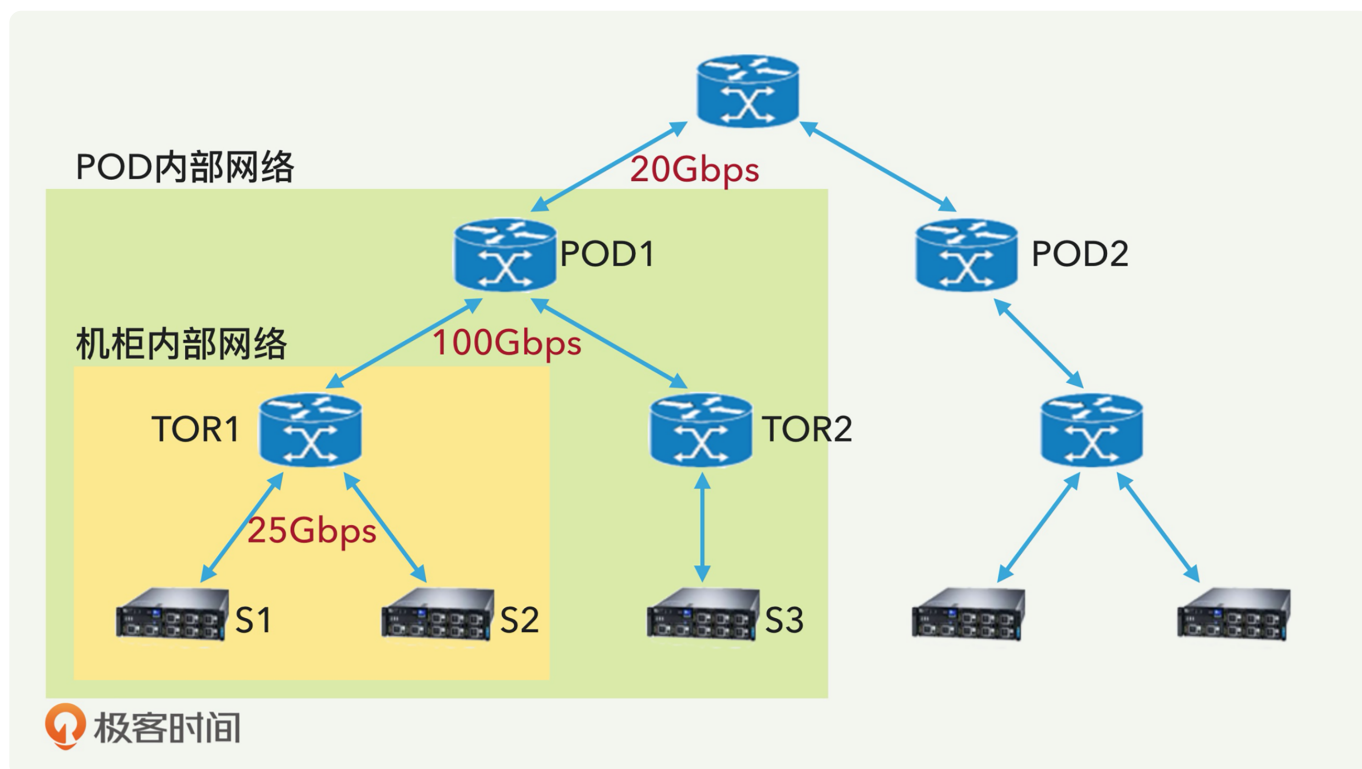
机柜交换器 (TOR, Top Of Rack; or RSW, Rack Switch)

数据中心里面的服务器不是单独放置的，一般是几十台服务器组成了一个机柜。机柜上面会有机柜交换器。这个机柜交换器的作用，一方面让机柜内部的服务器直接互通；另一方面，机柜交换器会有外联线路，连接到数据中心的骨干网络。

数据中心网络

数据中心网络里面也分了好几层，从 TOR 到集群交换器（Cluster Switch），再到集合交换器（Aggregation Switch）等，最后到数据中心路由器。

下面的图示，简单展示了机柜内部的网络和 POD 内部网络。TOR1 是机柜交换器，负责机柜内部几十台服务器之间的数据交换。POD1 内部网络包含很多机柜。



这里有一个值得注意的地方是，就是这个多层次结构，一般是越往上层，总的带宽越少。

比如，服务器的网卡带宽是 25Gbps，即使这个机柜内有 30 台服务器，总带宽就是 750Gbps，机柜交换器 TOR 的外联带宽有多少呢？可能只有 100Gbps。这个差距就叫**带宽超订**（over-subscription）。同样的，POD 交换器之间的带宽会继续变小。之所以允许带宽超订，是因为多数的数据交换是在内部进行的，不会全部都和外部进行交换。

知道这一点是必要的，因为我们做网络方面和服务部署优化的时候，需要考虑这点，不要让高层的网络带宽成为性能瓶颈。比如假设两个服务分别用不同的服务器，它们之间的数据交换如果很多的话，就尽量让它们运行在同一个机柜的服务器里面；如果不能保证同一个机柜，就尽量是同一个 POD 内部。

互联网的网络性能

讲完了数据中心内部的网络，我们继续往外扩展，讨论互联网的网络性能。互联网上运行的是 TCP/IP 协议。这两个协议本身比较复杂，有很多和性能相关的特性值得仔细学习。

一个常见的性能问题就是**丢包**。TCP 对丢包非常敏感，因为每次丢包，TCP 都认为是网络发生了拥塞，因此就会降低传输速度，并且采取重传来恢复，这就影响网络性能。

实际情况中，造成丢包的原因有很多，不一定就是网络拥塞。因此我们需要进行各种测试观测来做根因分析。

通常的丢包原因，是**端到端的网络传输中的某一段发生了问题**，或许是拥塞，或许是硬件问题，也或许是其他软件原因。我们需要一步步地逐段逐层地排除。

对于网络的每段，可以用工具（比如 Traceroute）来发现每一段路由，然后逐段测试。

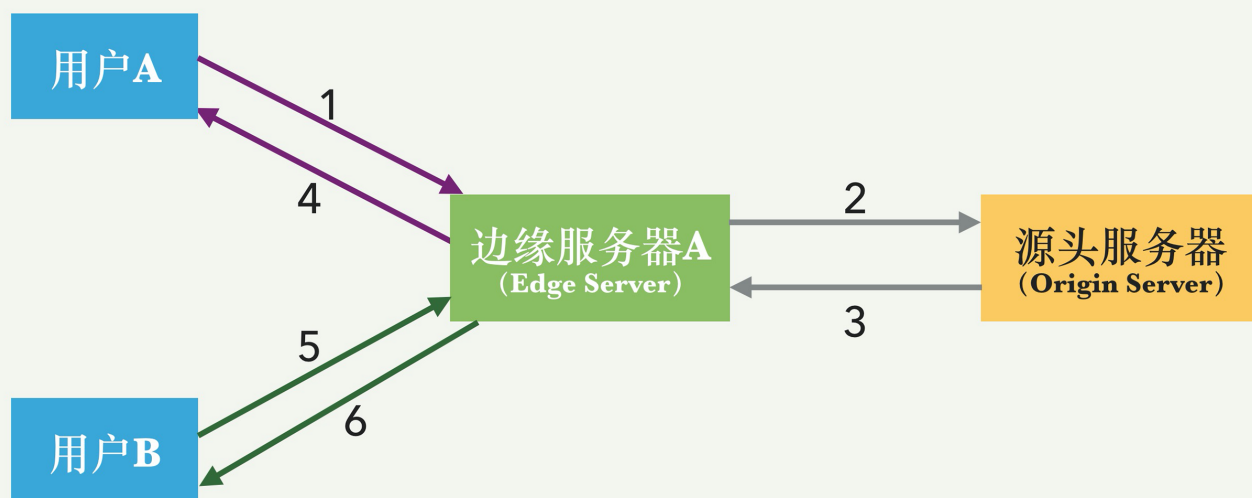
对于协议的每层，你都可以用相关工具进行分析。比如，你可以分为 TCP 层、操作系统、网卡驱动层，分别分析。具体来讲，对 TCP 层，可以用比如 netstat 来观察是不是套接字缓存不够；对操作系统，可以观察 softnet_stat，来判断 CPU 的查询队列；对网卡驱动层，可以用 ethtool 等来进行分析。

内容分发网络（CDN）的性能

当今互联网几乎普遍采用内容分发网络来提高网络性能。内容分发网络，也叫 CDN（Content Delivery Network 或 Content Distribution Network，CDN），是一种分布式网络，它可以有效地将 Web 内容交付给用户。

内容分发网络的基本原理是，利用最靠近每位用户的服务器，更快、更可靠地将（音乐、图片及其他）文件发送给终端用户，而不是每次都依赖于中心服务器。靠近用户的服务器，一般叫边缘服务器，会把请求的内容最大限度地缓存，以尽量地减少延迟。

我们用下图讲述它的基本工作方式，图里有两个用户和一个边缘服务器（Edge Server），以及一个源头服务器（Origin Server）。



第 1 步：用户 A 通过使用具有特殊域名的 URL 来请求文件。DNS 将请求重定向到性能最佳的边缘服务器，该位置通常是地理位置上最接近用户的服务器。

第 2 步：如果重定向的边缘服务器 A 中没有文件，则边缘服务器会向源头服务器请求该文件。源头服务器可以是任何可公开或非公开访问的 Web 服务器。

第 3 步：源头服务器将文件返回到边缘服务器。

第 4 步：边缘服务器先缓存文件，并将文件返回给原始请求者。该文件将保留在边缘服务器上，这样下次这个文件就可以迅速返回请求的客户。这个文件会保存到什么时候呢？时间由其 HTTP 标头指定的生存时间（TTL）到期为止。

第 5 步：如果用户 B 用相同 URL 请求相同的文件，也可能定向到相同的边缘服务器。

第 6 步：如果文件的 TTL 尚未过期，则边缘服务器直接从缓存中返回文件。这样的用户体验就更快。

内容分发网络的边缘服务器节点会在多个地点，多个不同的网络上摆放。这些节点之间通常会互相传输内容，对用户的下载行为最优化，并借此改善用户的下载速度，提高系统的稳定性。同时，将边缘服务器放到不同地点，也可以减少网络互连的流量，进而降低带宽成本。

内容分发网络提供商往往有很大的规模，比如几十万台服务器。对服务的客户提供所需要的节点数量会随着需求而不同。

工具

我们最后来看看网络性能有关的工具。这方面的性能测试和观测工具也不少，这里简单介绍几个。

Netperf 是一个很有用的网络性能的测量工具，主要针对基于 TCP 或 UDP 的传输。

Netperf 有两种操作模式：批量数据传输和请求 / 应答模式，根据应用的不同，可以进行不同模式的网络性能测试，Netperf 测试结果所反映的，是一个端到端的系统能够以多快的速度发送数据和接收数据。

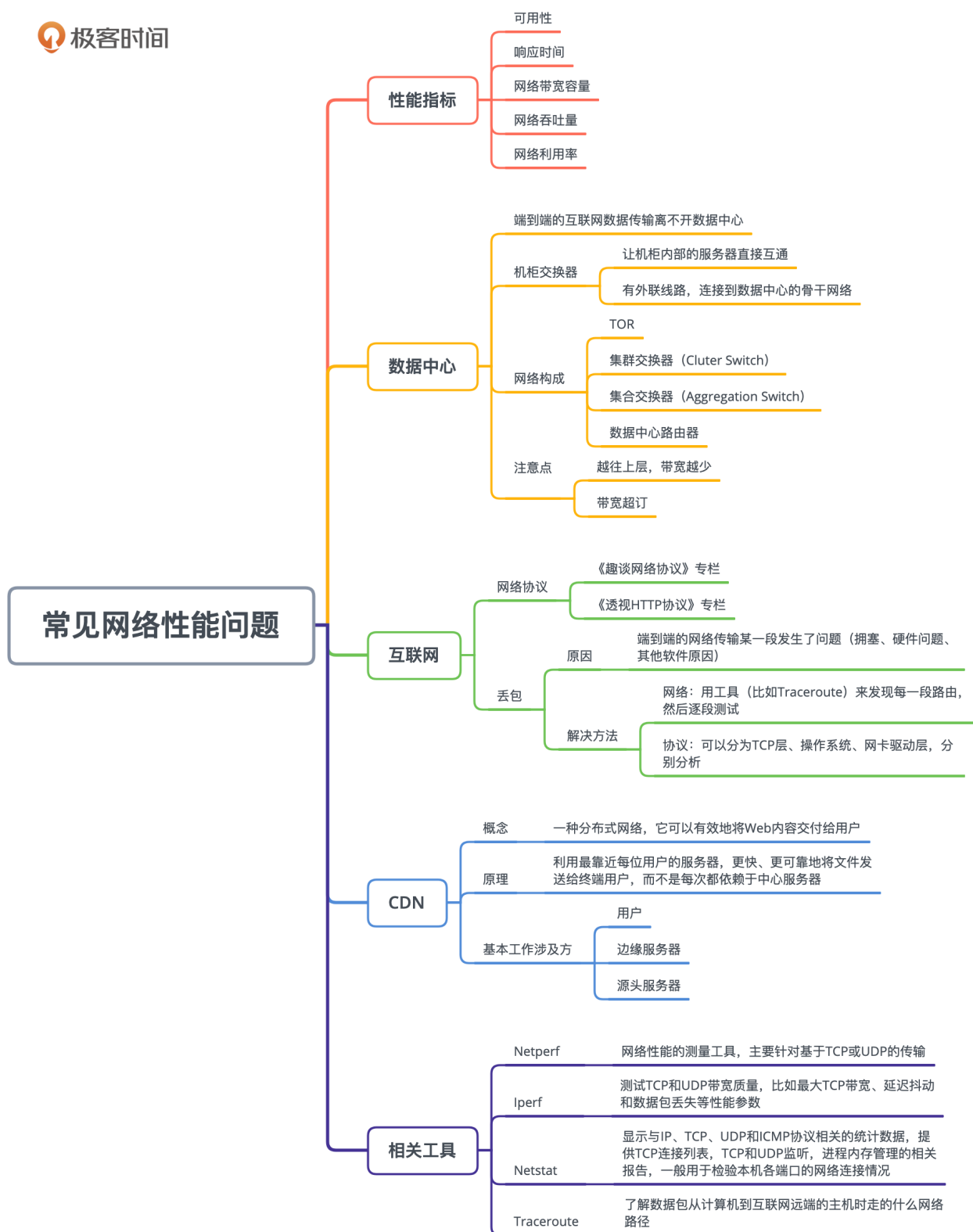
Iperf 这个工具也以测试 TCP 和 UDP 带宽质量，比如最大 TCP 带宽，延迟抖动和数据包丢失等性能参数。

Netstat 这一命令可以显示与 IP、TCP、UDP 和 ICMP 协议相关的统计数据，提供 TCP 连接列表，TCP 和 UDP 监听，进程内存管理的相关报告，一般用于检验本机各端口的网络连接情况。

Traceroute 这一命令可以帮我们知道，数据包从我们的计算机到互联网远端的主机，是走的什么网络路径。

总结

不管计算机技术如何发展，网络永远是极其重要的一部分；而且随着互联网业务的多样化和复杂化，对网络性能的要求只会越来越高。不管应用如何变化，高性能网络必不可少，而且会变得更复杂。这让我想起了宋朝大词人张先的《千秋岁》里面的几句词，“天不老，情难绝。心似双丝网，中有千千结。”



网络是整个互联网服务的一部分，各种因素之间互相影响。和网络性能演化密切相关的因素包括：上层服务需求的越来越多样化，计算机硬件和操作系统的持续进化，各层网络协议的不断优化，以及底层介质的演化（比如 5G）。

复杂的网络性能问题会受所有这些因素的影响，似有“千千结”。所以，网络性能分析和优化，需要仔细分析和考虑这些方面。

我们这一讲讨论了网络性能的几个性能指标，也讨论了单机网络协议，数据中心的网络，互联网和内容分发网络。

对于网络性能要求，简单来说就是能联通、响应快、带宽高；并且在部署大流量服务时，可以优化服务器的部署，来尽量减少外部网络流量；也可以使用 CDN 来加速数据传输。

思考题

你们公司的互联网服务使用内容分发网络，也就是 CDN 吗？是哪个 CDN 公司提供的？

除了你们使用的这家 CDN 公司，市场上还有哪些 CDN 提供商？选择 CDN 提供商的时候，你们会考虑哪些方面吗？比如规模、带宽、价格、服务质量等等。

欢迎你在留言区分享自己的思考，与我和其他同学一起讨论，也欢迎你把文章分享给自己的朋友。

点击查看 

21天打卡, Facebook 工程师 带你攻克性能工程



PC 端用户扫码参与



新版升级：点击「 请朋友读」，20位好友免费读，邀请订阅更有**现金**奖励。

上一篇 17 | 存储篇：如何根据性能优缺点选择最合适的存储系统？

下一篇 19 | 性能优化六大原则：三要三不要，快速有效地进行优化

精选留言 (1)

写留言



Scott

2020-01-07

这里POD网络的明确定义什么，去google的结果都是k8s的内容，肯定不是同一个东西

作者回复: POD是介于Rack（机架）和DC（数据中心）之间的逻辑单位。不同数据中心的设计可能不太一样，但意思是一样的。就是DC里面的一个层级，因为一个DC里面可能有几百个机架，所以需要再加一层，方便组织；尤其是网络的组织。

